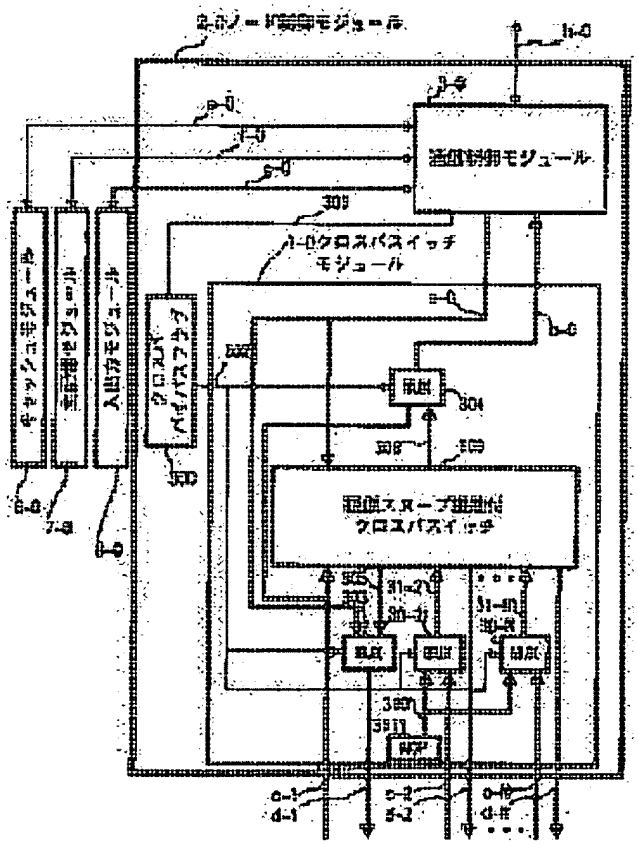
MAIN STORAGE SHARING TYPE MULTIPROCESSOR SYSTEM

Publication number: JP2000348000 (A)
Publication date: 2000-12-15
Inventor(s): TANAKA TAKESHI; AKASHI HIDEYA; TSUSHIMA YUJI; UEHARA KEITARO; HAMANAKA NAOKI; SHONAI TORU
Applicant(s): HITACHI LTD
Classification:
- international: ***G06F15/173; G06F12/00; G06F12/08; G06F15/16; G06F15/177;*** **G06F15/16; G06F12/00; G06F12/08;** (IPC1-7): G06F15/16; G06F15/173; G06F15/177
- European: G06F12/08B4N
Application number: JP19990156560 19990603
Priority number(s): JP19990156560 19990603

Also published as:
JP3721283 (B2)
US6789173 (B1)

Abstract of **JP 2000348000 (A)**

PROBLEM TO BE SOLVED: To reduce the cost of development for a main storage sharing type multiprocessor system by using a common node between a small scale system of a small number of nodes and a large scale system of a large number of nodes. SOLUTION: Each of nodes constructing a multiprocessor has a constitution shown in a diagram and also has plural processors having caches which are connected to a bus (h). A crossbar switch module 1-0 of a node control module 2-0 has a crossbar switch 100 having a pseudo snoop function having a cache coherency control function. A crossbar bypass flag 300 controls an MUX and transfers the information to other nodes other than its own one with no intervention of the switch 100 when the flag is set at 1 to invalidate the switch 100.; Then the information is transferred to other nodes via the switch 100 when the flag is set at 0 to validate the switch 100. An outside crossbar switch is placed outside the nodes and all node flags are set at 1 when the nodes are connected to the outside crossbar switch.



Family list
2 application(s) for: **JP2000348000 (A)**

1 **MAIN STORAGE SHARING TYPE MULTIPROCESSOR SYSTEM**

Inventor: TANAKA TAKESHI ; AKASHI HIDEYA (+4) **Applicant:** HITACHI LTD

EC: G06F12/08B4N **IPC:** ***G06F15/173; G06F12/00; G06F12/08;*** (+8)

Publication info: **JP2000348000 (A)** — 2000-12-15
JP3721283 (B2) — 2005-11-30

2 **Node controller for performing cache coherence control and memory-shared multiprocessor system**

Inventor: TANAKA TSUYOSHI [JP] ; AKASHI HIDEYA [JP] (+4) **Applicant:** HITACHI LTD [JP]

EC: G06F12/08B4N **IPC:** ***G06F15/173; G06F12/00; G06F12/08;*** (+6)

Publication info: **US6789173 (B1)** — 2004-09-07

Data supplied from the ***esp@cenet*** database — Worldwide

(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（Ａ）**　(11)特許出願公開番号
特開2000－348000
（P2000－348000A）
(43)公開日　平成12年12月15日(2000.12.15)

(51)Int.Cl.⁷		識別記号	FI			テーマコード（参考）
Ｇ０６Ｆ	15/16	６４５	Ｃ０６Ｆ	15/16	６４５	５Ｂ０４５
	15/173			15/173	Ｃ	
	15/177	６８２		15/177	６８２Ｊ	

審査請求　未請求　請求項の数5　ＯＬ　（全 12 頁）

(21)出願番号　特願平11－156560

(22)出願日　平成11年６月３日(1999.6.3)

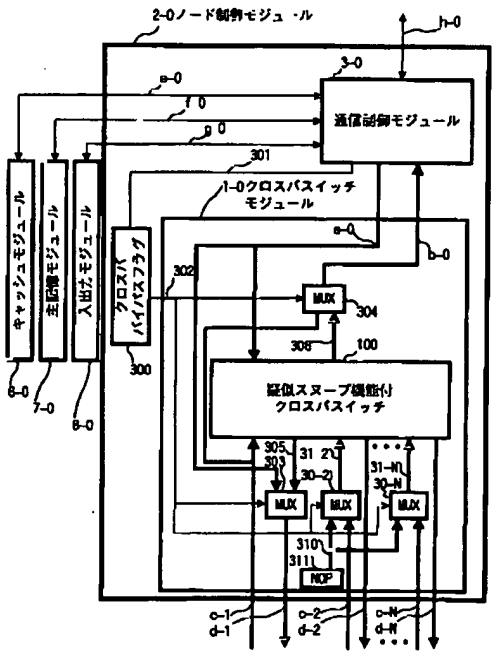
(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地
(72)発明者　田中　剛
東京都国分寺市東恋ケ窪一丁目280番地
株式会社日立製作所中央研究所内
(72)発明者　明石　英也
東京都国分寺市東恋ケ窪一丁目280番地
株式会社日立製作所中央研究所内
(74)代理人　100099298
弁理士　伊藤　修　（外１名）

最終頁に続く

(54)【発明の名称】　主記憶共有型マルチプロセッサシステム

(57)【要約】
【課題】　ノード数の少ない小規模システムとノード数の多い大規模システムで共通のノードを使うことにより開発コストを削減することにある。
【解決手段】　マルチプロセッサを構成する各ノードは図に示す構成にさらにバスｈにキャッシュを有するプロセッサを複数台接続した構成を有する。ノード制御モジュール2-0のクロスバスイッチモジュール1-0はキャッシュコヒーレンシ制御機能を持つ疑似スヌープ機能付クロスバスイッチ100を有し、クロスババイパスフラグ300はＭＵＸを制御し、フラグが１のとき他のノード間との情報転送は該スイッチ100を介さず行われ、該スイッチ100は無効になり、フラグが０のとき他のノード間との情報転送は該スイッチ100を介して行われ、該スイッチ100は有効となる。ノード外部に外部クロスバスイッチを設け、各ノードを該スイッチに接続する場合には、全ノードのフラグを全て１にする。

図3



【特許請求の範囲】
【請求項１】　複数のノードを有する主記憶共有型マルチプロセッサシステムにおいて、
該各ノードは、キャッシュメモリを有する１個以上のＣＰＵモジュールと、１個以上の主記憶モジュールと、これらのモジュールと他ノードの間の通信制御を行うノード制御モジュールを備え、
該ノード制御モジュールは、ノード間通信のインタフェースを制御する通信制御モジュールと、全ノードで発行するメモリアクセス要求の処理順序を決定してキャッシュコヒーレンス制御を行うクロスバスイッチを有するクロスバスイッチモジュールを有し、
該クロスバスイッチモジュールは、モードレジスタを有し、該モードレジスタにセットされた値に応じて該クロスバスイッチを有効、あるいは無効にすることを特徴とする主記憶共有型マルチプロセッサシステム。
【請求項２】　請求項１記載の主記憶共有型マルチプロセッサシステムにおいて、
前記クロスバスイッチモジュールは、前記モードレジスタに代えてモード信号ピンを有し、該モード信号ピンの信号値に応じて該クロスバスイッチを有効、あるいは無効にすることを特徴とする主記憶共有型マルチプロセッサシステム。
【請求項３】　請求項１または請求項２記載の主記憶共有型マルチプロセッサシステムにおいて、
前記クロスバスイッチモジュールは、全ノードから前記クロスバスイッチに転送されるメモリアクセス要求がどのノード内の主記憶モジュールに割り付けられているアドレス空間に対する要求であるかを判別する手段を有することを特徴とする主記憶共有型マルチプロセッサシステム。
【請求項４】　請求項１乃至請求項３のいずれかの請求項記載の主記憶共有型マルチプロセッサシステムにおいて、
前記ノードの外部に、全ノードで発行するメモリアクセス要求の処理順序を決定してキャッシュコヒーレンス制御を行うクロスバスイッチを有する外部クロスバスイッチモジュールを設け、
前記各ノードを該外部クロスバスイッチモジュールに接続し、
前記各ノード内のクロスバスイッチを無効にすることを特徴とする主記憶共有型マルチプロセッサシステム。
【請求項５】　請求項１乃至請求項３のいずれかの請求項記載の主記憶共有型マルチプロセッサシステムにおいて、
前記ノード間を直接接続し、少なくとも１つのノード内のクロスバスイッチを有効にすることを特徴とする主記憶共有型マルチプロセッサシステム。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、一般にプロセッサが主記憶を共有し、キャッシュメモリのコヒーレンシ制御のために、要求キャッシュラインのアドレスを全プロセッサに配布するようなスヌープ方式を使用しているマルチプロセッサシステムに係り、特に、ＣＰＵと主記憶、及びキャッシュメモリユニットが搭載され、小規模システムと大規模システムで共通の部品として使用できる構成のノードを有するマルチプロセッサシステムに関する。
【０００２】
【従来の技術】従来技術として、ＣＰＵモジュールと主記憶モジュールを同一のボードに搭載したノードを、小規模マルチプロセッサシステムと大規模マルチプロセッサシステムで共通に使用している装置については、「James Laudon, et. al. System Overview of the SGI Origin200／2000 Product Line. Proceeding ofthe 47th IEEE COMPUTER SOCIETY INTERNATIONAL CONFERENCE, pp. 150-156, Feb. 1997」に記載されている。Origin200／2000は、１枚以上のノードから構成され、各ノードにはＣＰＵ２個、主記憶、ディレクトリ、Hub chipから構成される。Hub chipは、ＣＰＵとの通信インタフェース制御部、主記憶及びディレクトリの通信インタフェース制御部、外部入出力インタフェース制御部、及びこれらのインタフェース制御部を結合するクロスバスイッチで構成される。小規模マルチプロセッサシステムに該当するOrigin200は、通常１枚のノードで構成されるが、２枚のノードをHub chipからの外部入出力インタフェースで直接接続した構成もある。大規模マルチプロセッサシステムに相当するOrigin2000では、２枚以上のノードボードをクロスバスイッチを搭載したルーターボードで接続している。以下の記述ではOrigin200とOrigin2000は区別せずOriginと呼ぶことにする。Originの例のように、システムの規模に関わらず同一のノードを使用して多様なシステム構成が可能なことは、開発コストの削減、開発期間の短縮に有効な手段である。
【０００３】また、OriginはccNUMA（cache-coherent Non-Uniform Memory Access）方式のマルチプロセッサで、ディレクトリ方式のキャッシュコヒーレンシ制御を行っている。この制御についての詳細は、「James Laudon, et. al. The SGI Origin:A ccNUMA Highly Scalable Server. Proceeding of 24th Annual Symposium on Computer Archtecture, pp. 2

41-251, Jun. 1997」に記載されている。Originのメモリアクセスは通常以下のように行われる。CPUで発行されたメモリアクセス要求は、要求するアドレスに対応する主記憶の存在するノードに転送され、該ノード内のディレクトリを検索する。ディレクトリは、要求したアドレスに対応するキャッシュラインごとに設けられ、どのノードのキャッシュメモリに転送され、どのような状態になっているか記録されている。ディレクトリを検索した結果、判明したノードのキャッシュメモリ、あるいは主記憶から読み出したデータは、要求発行元のCPUに転送される。

【0004】さらに、従来技術として、特開平9-138782に、キャッシュコヒーレンシ制御の為に、メモリアクセス要求の処理順序を決定する調停回路を有するクロスバスイッチが示されている。一般に、クロスバスイッチは、データの並列転送を持ち、スループット性能がバスより高いことは一般に知られている。しかし、メモリアクセスの順番が部分的に逆転し、キャッシュコヒーレンシを崩す可能性がある。Originではディレクトリ方式でキャッシュコヒーレンシを維持しているが、特開平9-138782で示されているマルチプロセッサシステムでは、CPUから発行されたメモリアクセス要求を、クロスバスイッチモジュール内の論理的に唯一の動作をする調停回路で一意の順番付けを行うことで、キャッシュコヒーレンシ制御を行っている。順番付けをされた要求は、クロスバスイッチモジュール内の選択回路を通して、CPUモジュールや主記憶モジュール、及び入出力モジュールに転送される。このように、クロスバスイッチにメモリアクセス要求の順番付け機能をもたせ、全てのCPUモジュールにメモリアクセス要求をブロードキャストすることでスヌープキャッシュを実現する方式を、以下においては擬似スヌープ方式と呼ぶことにする。また、擬似スヌープ方式を実現するためのメモリアクセス要求の順番付け機能を搭載したクロスバスイッチモジュールを擬似スヌープ機能付クロスバスイッチモジュールと呼ぶことにする。しかしながら、特開平9-138782のマルチプロセッサシステムでは、Originのように、大規模システムに用いられているノードだけを少数使用し、これらのノード間を直結して小規模システムを構成できるような装置にはなっていない。

【0005】

【発明が解決しようとする課題】上述のディレクトリ方式のキャッシュコヒーレンシ制御を行うマルチプロセッサシステムの場合、一般に、ディレクトリを参照を行う分だけLSI間転送の回数が多くなり、メモリレイテンシが増大するという課題がある。また、主記憶の量を増大する場合、ディレクトリの量も増大する。したがって、大容量主記憶を搭載する場合、大量のディレクトリ用のメモリが必要となるためコストが高くなる課題もある。本発明の目的は、上述のOriginと同様に、ノード数の少ない小規模システムとノード数の多い大規模システムで共通のノードを使うことによる開発コストを削減できるマルチプロセッサシステムを提供することにある。本発明の他の目的は、小規模システムでは複数ノードを直結してシステムを構成し、ノード外部のクロスバスイッチを省略できるマルチプロセッサシステムを提供することにある。さらに、本発明の他の目的は、ディレクトリを用いないキャッシュコヒーレンシ制御方式を使うことで、ディレクトリ参照のオーバヘッドによるメモリレイテンシの増分を削減し、かつ主記憶を増加させても、主記憶以外のコストが増加しないマルチプロセッサシステムを提供することにある。

【0006】

【課題を解決するための手段】上記目的を達成するため、本発明は、複数のノードを有する主記憶共有型マルチプロセッサシステムにおいて、該各ノードは、キャッシュメモリを有する1個以上のCPUモジュールと、1個以上の主記憶モジュールと、これらのモジュールと他ノードの間の通信制御を行うノード制御モジュールを備え、該ノード制御モジュールは、ノード間通信のインタフェースを制御する通信制御モジュールと、全ノードで発行するメモリアクセス要求の処理順序を決定してキャッシュコヒーレンス制御を行うクロスバスイッチを有するクロスバスイッチモジュールを有し、該クロスバスイッチモジュールは、モードレジスタを有し、該モードレジスタにセットされた値に応じて該クロスバスイッチを有効、あるいは無効にするようにしている。

【0007】また、前記クロスバスイッチモジュールは、前記モードレジスタに代えてモード信号ピンを有し、該モード信号ピンの信号値に応じて該クロスバスイッチを有効、あるいは無効にするようにしている。

【0008】また、前記クロスバスイッチモジュールは、全ノードから前記クロスバスイッチに転送されるメモリアクセス要求がどのノード内の主記憶モジュールに割り付けられているアドレス空間に対する要求であるかを判別する手段を有するようにしている。

【0009】また、前記ノードの外部に、全ノードで発行するメモリアクセス要求の処理順序を決定してキャッシュコヒーレンス制御を行うクロスバスイッチを有する外部クロスバスイッチモジュールを設け、前記各ノードを該外部クロスバスイッチモジュールに接続し、前記各ノード内のクロスバスイッチを無効にするようにしている。

【0010】また、前記ノード間を直接接続し、少なくとも1つのノード内のクロスバスイッチを有効にするようにしている。

【0011】

【発明の実施の形態】以下、本発明に係わるマルチプロセッサを図面に示したいくつかの実施例を参照してさら

に詳細に説明する。

《実施例1》

（装置構成の概略）図1は、本発明に係わる主記憶共有型マルチプロセッサシステムの装置構成の概略を示す。図1において、5－i（0≦i≦N、Nは1以上の自然数）はノードであり、複数のノード間を信号線c－i、d－iで結合する。次にノード5－iの構成について説明する。各ノードは全て同一構造をしている。ノード5－iは、CPUモジュール4－i－j（1≦j≦k、kは1以上の自然数）、キャッシュモジュール6－i、主記憶モジュール7－i、入出力モジュール8－i、ノード制御モジュール2－iから構成される。ノード制御モジュール2－iは、通信制御モジュール3－i、クロスバスイッチモジュール1－iから構成されている。

【0012】各CPUモジュール4－i－jは、ライトバック方式のプロセッサキャッシュ（図示せず）を有する。また、各CPUモジュールはプロセッサバスh－iで結合されている。キャッシュモジュール6－iは、ライトバック方式のキャッシュメモリ（図示せず）とキャッシュメモリ制御部（図示せず）で構成されている。キャッシュモジュール6－iのキャッシュメモリは、ノード5－iの全CPUモジュール4－i－jで共有されている。本実施例のマルチプロセッサシステムのメモリ階層構造においては、プロセッサキャッシュと主記憶の中間に位置するキャッシュである。主記憶モジュール7－iは全ノード5－iで共有される主記憶空間の一部を構成している。入出力モジュール8－iは複数の入出力装置、例えば、ディスク装置9－iに接続されている。この入出力モジュールには、他の入出力装置、例えば、回線接続装置（図示せず）等も接続されている。

【0013】通信制御モジュール3－iは、CPUモジュール4－i－j、キャッシュモジュール6－i、主記憶モジュール7－i、及び入出力モジュール8－iと、それぞれ、プロセッサバスh－i、信号線e－i、f－i、g－iで結合している。さらに、通信制御モジュール3－iは、クロスバスイッチモジュール1－iと、信号線a－i、b－iで結合している。通信制御モジュール3－i（内部の構造は図示せず）は、各モジュール間の通信制御を行っている。例えば、CPUモジュール4－i－jからプロセッサバスh－iに転送されたメモリアクセス要求をキャッシュモジュール6－iに転送する処理や、他ノードに発行したメモリアクセス要求をクロスバスイッチモジュールを通して転送するための、通信プロトコル制御や、通信を行うデータのフォーマット変更等を行っている。クロスバスイッチモジュール1－iは、自ノード内の通信制御モジュール3－iと信号線a－i、b－iで結合し、他ノードと信号線c－i、d－iで結合している。

【0014】本発明では、同一構成のノードを、小規模構成のマルチプロセッサシステム、大規模構成のマルチプロセッサシステムで利用できることを特徴とする。ここで、マルチプロセッサシステムの規模とは、CPUモジュールの個数、つまり、ノードの枚数が少ないシステムを小規模、ノードの枚数が多いものを大規模と言っている。小規模、大規模の明確な数での切り分けはなく、実装上の都合でクロスバスイッチモジュール1－iに設けられている他ノードとの結合信号線c－i、d－iの組数によって決まる。図1は、小規模システムの構成を示し、図2は大規模システムの構成を示している。大規模構成システムの構成は、図1と同じ構成の複数ノード5－iを、ノード5－iの外に設けた外部クロスバスイッチモジュール20と、信号線c－0、…、c－N、d－0、…、d－Nで結合している。外部クロスバスイッチモジュール20は、前述した特開平9－138782号公報に記載されたものと同様のものである。

【0015】図3は、ノード制御モジュール2－0の内部を示している。ノードは全て同一構造なのでここでは図1のノード5－0を例にとり図3を用いて説明する。ノード制御モジュール2－0は、既に説明したとおり通信制御モジュール3－0と、クロスバスイッチモジュール1－0を有している。クロスバスイッチモジュール1－0は、擬似スヌープ機能付クロスバスイッチ100、このクロスバスイッチ100を使用する有効モードか使用しない無効モードかを設定するモードレジスタであるクロスババイパスフラグレジスタ300（説明及び図面においては、クロスババイパスフラグレジスタ300をクロスババイパスフラグ300とも記載する）、マルチプレクサ303、304、30－v（2≦v≦N）から構成されている。ただし、ノード数が2枚の場合は、マルチプレクサ30－2、…、30－N、信号線31－2、…、31－N、c－2、…、c－N、d－2、…、d－Nは使用されない。

【0016】クロスババイパスフラグレジスタ300は、1ビットのフラグレジスタで信号線301で通信制御モジュール3－0と接続、信号線302でマルチプレクサ303、304、30－v（2≦v≦N）と接続されている。マルチプレクサ304は、擬似クロスバからの信号306と別ノードからの信号線c－1を、クロスババイパスフラグからの信号線302を使ってどちらか一方を選択し、信号線b－0に出力している。マルチプレクサ303は、擬似クロスバからの信号305と通信制御モジュール3－0からの信号線a－0を、クロスババイパスフラグからの信号線302を使ってどちらか一方を選択し、信号線d－1に出力している。

【0017】クロスババイパスフラグ300が1のとき、マルチプレクサ303、304共に、それぞれ信号線a－0、c－1を選択し、擬似スヌープ機能付クロスバスイッチ100をバイパスする。クロスババイパスフラグ300が0のときは、マルチプレクサ303、304はそれぞれ、信号線305、306を選択する。

【００１８】マルチプレクサ３０－ｖは、クロスババイパスフラグが0のとき、他ノードからの信号線ｃ－ｖを選択し、信号線３１－ｖで擬似スヌープ機能付クロスバスイッチ１００に接続する。クロスババイパスフラグ３００が1のときは、信号線３１０を選択し、トランザクションが存在しないことを示すＮＯＰ信号３１１をマルチプレクサ３０－ｖへ転送する。例えばオールゼロの信号がトランザクションが無い状態を示すならば、該ＮＯＰ信号は配線時に固定値オールゼロに設定される。

【００１９】このように、クロスババイパスフラグ３００を0にした場合を「擬似スヌープ機能付クロスバスイッチ１００が有効」、1にした場合を「擬似スヌープ機能付クロスバスイッチ１００が無効」と呼ぶことにする。図3の擬似スヌープ機能付クロスバスイッチ１００が無効の場合は、通信制御モジュール３－0の信号線ａ－0、ｂ－0がノードの入出力として直接使われる。本実施例1のマルチプロセッサシステムにおいて、擬似スヌープ機能付クロスバスイッチ１００は、1枚のノードでのみ有効にし、他のノードは全て無効にする。図2の外部クロスバスイッチモジュール２０を使った大規模構成の場合には、図3の擬似スヌープ機能付クロスバスイッチ１００は、全ノードで無効にする。

【００２０】図3のクロスババイパスフラグレジスタ３００は、アドレス空間に配置したマップドレジスタにしておき、システム起動時に一枚のノードのクロスババイパスフラグレジスタだけ0にしておき、残りのノードのクロスババイパスフラグレジスタは全て1になるように設定する。または、図4のように、モード信号ピン（クロスババイパスモード信号ピン）４０を設置し、ボードの配線時にあらかじめ0か1に決めて配線してしまう方法や、ジャンパスイッチにしてあらかじめ手動で0か1に設定する方法（図示せず）を使用してもよい。

【００２１】（擬似スヌープ機構付クロスバスイッチの機能説明）擬似スヌープ機構付クロスバスイッチは、キャッシュコヒーレンシ制御のために、特開平９－１３８７８２に示されているメモリアクセス要求の順番付け機能を有している。クロスバスイッチは、データの並列転送ができることが特徴であることは一般に知られている。しかし、メモリアクセスの順番が部分的に逆転してしまい、キャッシュコヒーレンシを保証できない場合が起きうる。キャッシュコヒーレンシ制御方式としてバススヌープ方式がよく知られている。このバススヌープ方式では、バスに接続された複数のＣＰＵがＣＰＵバスにメモリアクセス要求を出すために、バス制御部にはバスの使用権を調停する機能がある。バスの調停で決まったメモリアクセス要求の順番が、システム全体で、つまりどのＣＰＵでも一意であることでキャッシュの一貫性を保っている。このような、システムで一意のメモリアクセス要求の順番を決定する機能をクロスバスイッチに搭載することで擬似スヌープ機構付クロスバスイッチはキャッシュコヒーレンシ制御を行っている。

【００２２】（装置動作）ここでは、ＣＰＵモジュールで発行されたメモリアクセス要求が、全ＣＰＵモジュール、及び主記憶モジュールにブロードキャストされる動作を説明する。なお、以降では、メモリアクセスの要求や、それに対する応答等、プロセッサバスやクロスバスイッチなどで転送されるデータをトランザクションと呼ぶことにする。

【００２３】図1のＣＰＵモジュール４－０－1において、ＣＰＵが内蔵のキャッシュメモリに所望のデータが存在せず、キャッシュミスを起こしたとする。この場合、ＣＰＵモジュール４－０－1はプロセッサバスｈ－0にキャッシュミスしたアドレスのデータをアクセスするトランザクションを発行する。該トランザクションは、メモリの読出しである等のトランザクションの種類の情報、トランザクションを発行したＣＰＵモジュール番号等の制御情報、要求するアドレスで構成されている（図示せず）。プロセッサバスｈ－0に転送されたトランザクションは、ノード制御モジュール２－0内の通信制御モジュール３－0を通して、キャッシュモジュール６－0に転送する。このキャッシュモジュール６－0で、該トランザクションの要求するアドレスのデータが保持されていない場合、所望のデータが他のノードでキャッシュされているかどうか調べるために全ノードにこのトランザクションを送ろうとする。この場合、通信制御モジュール３－0を通して、クロスバスイッチモジュール１－0に転送される。

【００２４】本動作例では、ノード５－0のクロスババイパスフラグ３００が0でクロスバスイッチが有効となっている。他のノード５－ｉ（１≦ｉ≦Ｎ）ではクロスババイパスフラグ３００が1となりクロスバスイッチが無効となっている。該クロスバスイッチモジュール１－0に入力されたトランザクションは信号線ａ－0を通して擬似スヌープ機能付クロスバスイッチ１００に転送される。

【００２５】図5は、図3に示されている擬似スヌープ機能付クロスバスイッチ１００の構成を示している。入力信号線ａ－0、ｃ－1、３１－2、…、３１－Ｎは、入力レジスタ５００－ｘ（0≦ｘ≦Ｎ）、５２０－ｘ、５２１－ｘに接続している。ただし、図5では、全ての信号の図示はしていない、代表してａ－0と３１－Ｎのみを示している。ＴＹＰＥレジスタ５００－ｘには、クロスバスイッチに転送されるメモリアクセス要求、ならびにその応答データなど、トランザクションの種類を表す情報が格納される。ＭＩＳＣレジスタ５２０－ｘには、どこのノードで発行されたトランザクションであるか等、トランザクションの制御情報が格納される。ＡＤＤＲＥＳＳレジスタ５２１－ｘには、要求しているアドレス情報が格納される。また、トランザクション全体は、信号線５０３－0を通し、トランザクションレジス

タ５０７－０に格納され、信号線ｔ０を通して選択回路５１２－ｘに転送され、調停完了を待つ。

【００２６】トランザクションの種類を表すＴＹＰＥレジスタ５００－０のデータは信号線５０２－０を通してリクエスト制御５０４－０へ転送される。そして、レジスタ５０６－０の中の転送対象のノードに対応する要素に調停要求信号を書き込む。ここでは、キャッシュを持ったモジュールがあるノード、あるいは、主記憶が転送先となるので全ノードが転送対象となる。このレジスタ５０６－０から調停要求信号ｒ００、…、ｒ０Ｎが調停回路５１０－０、…、５１０－Ｎに転送される。

【００２７】全ての調停回路が同じ調停動作を同時に行っている場合は、誤った順番付けを行うことはない。しかし、何らかの理由で出力レジスタ５１４－ｘ、５１５－ｘ，５１６－ｘにセットできず、各調停回路で調停を行う時間が異なる場合がありうるが、全調停回路は全て同じ優先順位で調停を行っているので、全て同じ順序でトランザクションが出力されていけばキャッシュコヒーレンシが崩れることはない。

【００２８】調停回路５１０－ｘ（０≦ｘ≦Ｎ）で調停に勝利した場合、調停完了信号ｇ０ｘがリクエスト制御５０４－ｘに転送される。どのノードからのトランザクションが調停で勝ったかという情報は、信号線５１１－ｘを通し選択回路５１２－ｘに伝えられる。いま、調停要求ｒ００が調停に勝利したとする。調停回路５１０－ｘは、リクエスト制御５０４－０に調停完了信号ｇ０ｘを送り、同時に、選択回路５１２－０に信号線５１１－０を通して通知する。選択回路５１２－ｘは、調停に勝ったノードを選択し、トランザクションを信号線ｔ０、及び信号線５１３－ｘを通し、出力レジスタ５１４－ｘ、５１５－ｘ、５１６－ｘに転送する。リクエスト制御５０４－ｘは、調停完了信号ｇ０ｘが到着した後、調停要求の入ったレジスタ５０６－０をリセットし、同時に信号線５０５－０を通してトランザクションレジスタ５０５－０から削除を指示する。調停要求コマンドを格納するレジスタ５０６－０、及びトランザクションレジスタ５０５－０はリセットされ、次の入力を待つ。出力レジスタのトランザクションは、図５の信号線３０６、…、ｄ－Ｎ（正しくは、図３の３０６、３０５、ｄ－２、…、ｄ－Ｎ）を通して各ノードに転送される。

【００２９】図１において、調停後のトランザクションは、ノード５－０ではクロスバスイッチモジュール１－０から信号線ｂ－０を通し通信制御モジュールを通過し、キャッシュモジュール６－０、プロセッサバスｈ－０に転送され、各々のキャッシュタグを検索する。ノード５－１、…、５－Ｎでは、入力されたトランザクションは、クロスバスイッチをバイパスして通信制御モジュールへ入力され、同様にキャッシュモジュール６－１、…、６－Ｎ、プロセッサバスｈ－１、…、ｈ－Ｎに転送され、各々のキャッシュタグを検索する。キャッシュの状態を調査した後、各々のノードはこの調査結果をトランザクション発行元に報告し（この手段は図示していないが専用線を設けたり、メモリアクセス要求の転送で用いたクロスバスイッチを使ってもかまわない）、主記憶、あるいはキャッシュメモリからデータがトランザクション発行元に転送され、メモリアクセス要求は完了する。このデータ転送手段は、クロスバスイッチモジュールを併用する場合もあるし、別のデータ専用の通信路を用意してもよい。

【００３０】《実施例２》

(装置構成の概略）図６は、本発明に係る主記憶共有型マルチプロセッサシステムの第２の実施例の装置構成の概略を示す。図６において、５－ｉ（０≦ｉ≦Ｎ、Ｎは１以上の自然数）はノードであり、複数のノード間を信号線６０－ｉｊ、（０≦ｊ≦Ｎ）で結合する。次にノード５－ｉの構成について説明する。各ノードは全て同一構造を有している。ノード５－ｉは、ＣＰＵモジュール４－ｉ－ｊ（１≦ｊ≦ｋ、ｋは１以上の自然数）、キャッシュモジュール６－ｉ、主記憶モジュール７－ｉ、入出力モジュール８－ｉ、ノード制御モジュール２－ｉから構成される。ノード制御モジュール２－ｉは、通信制御モジュール３－ｉ、クロスバスイッチモジュール１－ｉから構成されている。

【００３１】各ＣＰＵモジュール４－ｉ－ｊは、ライトバック方式のプロセッサキャッシュ（図示せず）を有する。また、各ＣＰＵモジュールはプロセッサバスｈ－ｉで結合されている。キャッシュモジュール６－ｉは、ライトバック方式のキャッシュメモリ（図示せず）とキャッシュメモリ制御部（図示せず）で構成されている。キャッシュモジュール６－ｉのキャッシュメモリは、ノード５－ｉの全ＣＰＵモジュール４－ｉ－ｊで共有されている。本実施例のマルチプロセッサシステムのメモリ階層構造においては、プロセッサキャッシュと主記憶の中間に位置するキャッシュである。主記憶モジュール７－ｉは全ノード５－ｉで共有される主記憶空間の一部を構成している。入出力モジュール８－ｉは複数の入出力装置、例えば、ディスク装置９－ｉに接続されている。この入出力モジュールには、他の入出力装置、例えば、回線接続装置（図示せず）等も接続されている。

【００３２】通信制御モジュール３－ｉは、ＣＰＵモジュール４－ｉ－ｊ、キャッシュモジュール６－ｉ、主記憶モジュール７－ｉ、入出力モジュール８－ｉを、それぞれ、プロセッサバスｈ－ｉ、信号線ｅ－ｉ、ｆ－ｉ、ｇ－ｉと結合し、クロスバスイッチモジュール１－ｉとは、信号線ａ－ｉ、ｂ－ｉで結合している。通信制御モジュール３－ｉ（内部の構造は図示しせず）は、各モジュール間の通信制御を行っている。通信制御の内容は、ＣＰＵモジュール４－ｉ－ｊからプロセッサバスｈ－ｉに転送されたメモリアクセス要求をキャッシュモジュール６－ｉに転送する処理や、他ノードに発行したメモリ

アクセス要求をクロスバスイッチモジュールを通して転送するための通信プロトコル制御や、通信を行うデータのフォーマット変更等である。

【００３３】クロスバスイッチモジュール１－ｉは、自ノード内の通信制御モジュール３－ｉと信号線ａ－ｉ、ｂ－ｉで結合し、他ノードと信号線６０－ｉｊで結合している。図７に示すアドレスマップ７００は、主記憶空間のアドレスと、どのノードの主記憶に割り付けられているかアドレスとノードの対応表である。アドレスマップ７００はメモリマップドレジスタとして主記憶空間に配置され、システムが立ち上がるとき、ノードに割り振られているメモリ空間のアドレスをセットする。また、ノード上のジャンパスイッチなどを使用してアドレスマップに格納するアドレス情報を定義する方法もある。

【００３４】図７は、ノード制御モジュール２－０の内部を示している。ノードは全て同一構造で、同一動作を行うので、本実施例では図６のノード５－０の動作のみを図７を使って説明する。ノード制御モジュールは、既に説明したとおり通信制御モジュール３－０と、クロスバスイッチモジュール１－０を有している。クロスバスイッチモジュール１－０は、擬似スヌープ機能付クロスバスイッチ１００、トランザクション判別部７０－０、アドレスマップ７００、マルチプレクサ７０４，７０５、クロスババイパスフラグレジスタ７１０（説明及び図面においては、クロスババイパスフラグレジスタ７１０をクロスババイパスフラグ７１０とも記載する）から構成されている。クロスババイパスフラグは、実施例１の図３に示されているものと同様の働きをし、図７のクロスババイパスフラグ７１０を１に設定すると、通信制御モジュール３－０と擬似スヌープ機能付クロスバスイッチ間の信号線ａ－０、ｂ－０がノードの入出力信号６０－０１、６０－１０と直結する。また、クロスババイパスフラグが０の時は、通信制御モジュールの信号線ａ－０はトランザクション判別部７０－０に、信号線ｂ－０は擬似スヌープ機能付クロスバスイッチ１００からの出力信号となる。

【００３５】クロスババイパスフラグは、ノード間を直結する小規模システム時には、全ノードで０に設定し、外部クロスバスイッチモジュールでノードを結合する大規模構成時には１に設定する。なお、クロスババイパスフラグ７１０は信号線７１１で通信制御モジュール３－０と接続し、実施例１と同様にメモリマップドレジスタにして、システム起動時に設定するようにする。あるいは、実施例１と同様にクロスババイパスフラグレジスタの代わりにモード信号ピン（図示せず）を設ける方法でもよい。

【００３６】トランザクション判別部７０－０は、通信制御モジュールからの出力信号線ａ－０とアドレスマップ７００からの信号線７０２を入力とし、疑似スヌープ機能付クロスバスイッチ１００と接続する信号線７１－０を出力する。擬似スヌープ機能付クロスバスイッチ１００の疑似スヌープ機能については実施例１で説明した通りである。図７のクロスバスイッチモジュール１－０内の疑似スヌープ機能付クロスバスイッチ１００は、図１２に示す表に基づいた動作を行う。図１２に基づく動作をフローチャートにしたものが図９である。

【００３７】図８において、トランザクション判別部７０－ｉは、入力レジスタ８０１、８０２、８０３、８０４、８０５、及び８０６、要求ノード検索回路８００から構成されている。入力レジスタは８０１、８０２、８０３、８０４、８０５、及び８０６は信号線ａ－ｉと接続している。要求ノード検索回路８００は入力として、入力レジスタ８０４と信号線８０７で、及びアドレスマップと信号線７０２で接続されている。要求ノード検索回路８００の出力信号で、ＡＭレジスタ８０９の内容を変更する。該ＡＭレジスタには、ノード番号が格納されている。このノード番号は、トランザクションが要求しているアドレスがどのノードの主記憶に割り付けられているか示している。

【００３８】（装置動作）ここでは、実施例１と同様にＣＰＵモジュールで発行されたメモリアクセス要求を行うトランザクションが、全ＣＰＵモジュール、及び主記憶モジュールにブロードキャストされる動作を説明する。本実施例では、全ノードの主記憶上のデータ、あるいは他ノードのキャッシュにアクセスを行う場合を説明する。本実施例では、メモリアクセス要求を行うトランザクション発行元が自ノードの主記憶に割り付けられたアドレスをアクセスする場合と他ノードの主記憶に割り付けられたアドレスをアクセスする場合を示す。

【００３９】まず、動作例（１）として自ノードの主記憶に割り付けられたアドレスをアクセスする場合について説明する。図６のＣＰＵモジュール４－０－１において、ＣＰＵが内蔵のキャッシュメモリに所望のデータが存在せず、キャッシュミスを起こしたとする。この場合、ＣＰＵモジュール４－０－１はプロセッサバスｈ－０にキャッシュミスしたアドレスのデータをアクセスするトランザクションを発行する。該トランザクションは、メモリの読出しである等のトランザクションの種類の情報、トランザクションを発行したＣＰＵモジュール番号等の制御情報、要求するアドレスで構成されている（図示せず）。プロセッサバスｈ－０に転送されたトランザクションは、ノード制御モジュール２－０内の通信制御モジュール３－０を通して、キャッシュモジュール６－０に転送する。該キャッシュモジュールで、当該トランザクションの要求するアドレスのデータが保持されていない場合、所望のデータが他のノードでキャッシュされているかどうか調べるために全ノードにこのトランザクションを送ろうとする。この場合、通信制御モジュール３－０から信号線ａ－０を通して、クロスバスイッチモジュール１－０に転送される。

【００４０】クロスバスイッチモジュール１－０に転送されたトランザクションは、図8のトランザクション判別部７０－ｉに転送される。まず、入力レジスタ８０１、８０２、８０３、８０４、８０５、８０６にトランザクションを格納する。ＴＹＰＥレジスタ８０２はトランザクションの種類を示す情報が格納される。ＭＩＳＣレジスタ８０３には、トランザクションのＩＤ番号等の制御情報が格納されている。ＡＤＤＲＥＳＳレジスタ８０４には、トランザクションのアクセスするアドレスが、ＮＩＤレジスタ８１３には、トランザクションを発行したノードの番号を示す情報が格納されている。また、ＡＭレジスタ８０６には、トランザクションが要求するアドレスが、どのノードの主記憶に割り付けられているか、そのノード番号が格納されている。ＡＲＢレジスタ８０１は、１ビットのレジスタで当該トランザクションがクロスバスイッチモジュールで既に調停を終了している場合は１、調停未完了のときは０が格納されている。本ケースでは、まだクロスバスイッチで調停が行われていないのでＡＲＢ＝０となっている。要求ノード検索回路８００は、アクセス対象が自ノードに設定されたメモリ空間なのでＡＭレジスタに自ノード番号である‘０’を書き込む。ただし、本実施例では、図6のノード５－ｉ（０≦ｉ≦Ｎ）のＮＩＤをｉとしている。

【００４１】図１０の疑似スヌープ機能付クロスバスイッチ１００に入力されたトランザクションは、入力レジスタであるＴＹＰＥレジスタ５００－０、ＭＩＳＣレジスタ５２０－０、ＡＤＤＲＥＳＳレジスタ５２１－０、ＡＲＢレジスタ１１００－０、ＮＩＤレジスタ１１０２－０、ＡＭレジスタ１１０３－０に格納される。レジスタ１１００－０，５００－０，５２０－０、１１０２－０、１１０３－０はコマンド信号としてリクエスト制御５０４－０に送られ、送付ノードの番号のついたレジスタ５０６－０に調停要求信号をセットする。また、トランザクション全体は信号線５０３－０を通してトランザクションレジスタ５０７－０に格納される。

【００４２】図9のフローチャートに示す判定をリクエスト制御５０４－ｉ（０≦ｉ≦Ｎ）で行う。本ケースでは、ＡＲＢ＝０，ＡＭ＝０なので図9の条件１００１は満たさず、条件１００３を満たすので、図9の処理１００４を実行する。処理１００４により、信号線５０５－０でＡＲＢ＝１にセットし、レジスタ５０６－０は全ノードを対象に調停要求をセットする。

【００４３】リクエスト制御５０４－０から調停回路５１０－ｉに転送された調停要求ｒ００、…、ｒ０Ｎが調停回路で勝利すると選択回路５１２－０に勝利したノードの番号、ここでは０を伝え、信号線ｔ０を通してトランザクションを出力レジスタ１１０１ｉ、５１４－ｉ、５１５－ｉ、５１６－ｉ、１００４－ｉ、１１０５ｉにセットし、各ノードに転送する。他ノードには、図7の信号線６０－０ｉ（１≦ｉ≦Ｎ）を通して伝えられる。自ノードの通信制御モジュール３－０にはｂ－０を通してトランザクションが伝えられる。

【００４４】トランザクションが他ノードに到着した場合を説明する、ここでは簡単の為、図6のノード５－１での場合を説明する。図6のノード５－０の信号線６０－０１を通してトランザクションがノード５－１に伝えられる。当該トランザクションは信号線６０－０１を通して図１１の疑似スヌープ機能付クロスバスイッチ１００に転送される。疑似スヌープ機能付クロスバスイッチ１００では、図9のフローチャートに従い、ＡＲＢ＝１なので条件１００１を満足するので１００２の処理を行う。疑似スヌープ機能付クロスバスイッチ１００の動作は先の説明とほとんど同様だが、図１０のリクエスト制御５０４－１のレジスタ５０６－１で自ノードに対応するレジスタのみ調停要求をセットする。後の処理は、先ほどと同様の調停処理を行い、図１１の信号線ｂ－１を通して通信制御モジュール３－１にトランザクションを転送する。このようにして、全ノードの通信制御モジュールにトランザクションがブロードキャストされる。ブロードキャスト後の処理は実施例１の動作例と同様である。

【００４５】次に、動作例（２）として他ノードの主記憶に割り付けられたアドレスをアクセスする場合について説明する。ここでは、図6のノード５－１に割り当てられた主記憶にアクセスすると仮定する。図6のＣＰＵモジュール４－０－１で発行されたトランザクションがキャッシュミスを起こし、クロスバスイッチモジュール１－０に転送されるまでの動作は動作例（１）の場合と同様である。

【００４６】図6のクロスバスイッチモジュールにトランザクションが転送された後、図１０の擬似スヌープ機能付クロスバスイッチのリクエスト制御５０４－０では、図9の処理が行われる。本動作例では、ＡＲＢ＝０、ＡＭ＝１なので、条件１００１、１００３は満たされず、処理１００５が行われる。処理１００５では、ＡＭレジスタ１１０３－０で示されたノードにのみトランザクションを転送するので、図１０のレジスタ５０６－０において、該ＡＭレジスタで示されているノードのみに調停要求をセットする。調停後、該トランザクションは、図6のノード５－１に転送される。当該ノードの内蔵する擬似スヌープ機能付クロスバスイッチでは、図9の１００１の条件は満たさず、条件１００３を満たすので処理１００４が行われる。この後の動作は動作例（１）と同様である。

【００４７】

【発明の効果】本発明によれば、各ノードにクロスバスイッチを搭載し、直接ノード間結合することで小規模マルチプロセッサシステムを構成できる。しかも、このノードは大規模システムにも利用できるため開発コストを削減できる。さらに、本発明によれば、ディレクトリを

用いずクロスバスイッチでキャッシュコヒーレンシ制御をするため、ディレクトリを検索するための通信コストが削減されメモリレイテンシの削減が可能となる。また、主記憶を増加する場合、ディレクトリを使用していないため主記憶以外の資源を追加する必要がなく、低コストでシステムのスケールアップが可能となる。

【図面の簡単な説明】

【図１】第１の実施例に係わる主記憶共有型マルチプロセッサシステムの装置構成の概略を示す図である。

【図２】第１の実施例におけるノードを外部クロスバスイッチモジュールに接続した大規模マルチプロセッサシステムの全体構成を示す図である。

【図３】第１の実施例におけるノード制御モジュールの構成を示す図である。

【図４】図３に示すノード制御モジュールの変形例を示す図である。

【図５】第１の実施例における疑似スヌープ機能付クロスバスイッチの構成を示す図である。

【図６】第２の実施例に係わる主記憶共有型マルチプロセッサシステムの装置構成の概略を示す図である。

【図７】第２の実施例におけるノード制御モジュールの構成を示す図である。

【図８】第２の実施例におけるトランザクション判別部の構成を示す図である。

【図９】第２の実施例における疑似スヌープ機能付クロスバスイッチを構成するリクエスト制御の動作の一部のフローチャートを示す図である。

【図１０】第２の実施例における疑似スヌープ機能付クロスバスイッチの構成を示す図である。

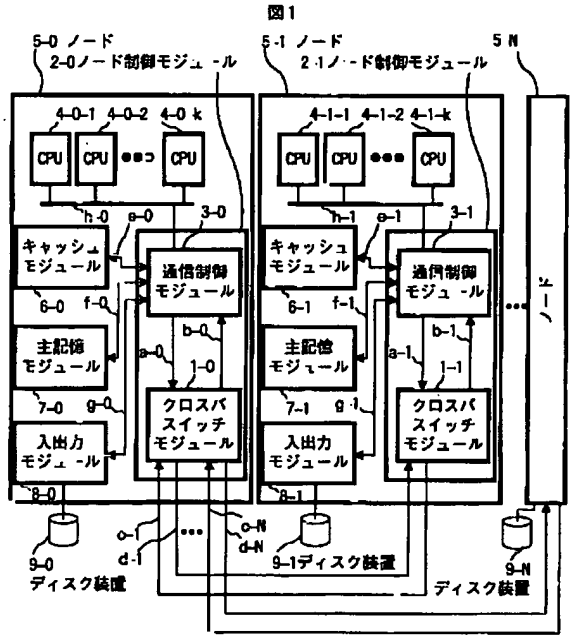
【図１１】第２の実施例における他のノードのノード制御モジュールの構成を示す図である。

【図１２】第２の実施例における疑似スヌープ機能付クロスバスイッチを構成するリクエスト制御の動作の一部を記述した表を示す図である。

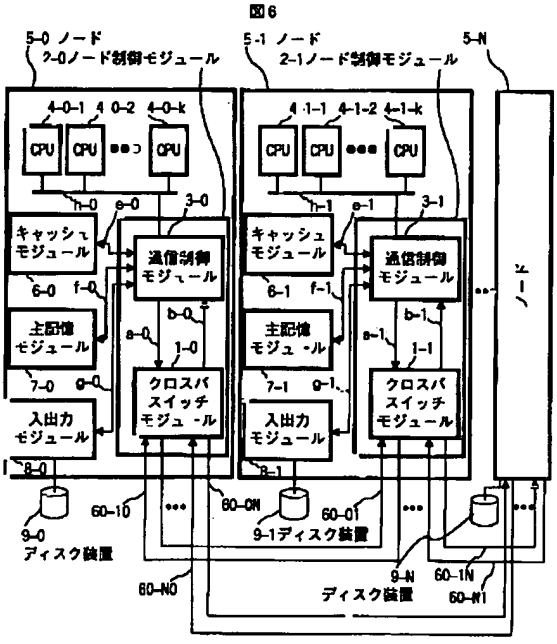
【符号の説明】

１－０，１－１　クロスバスイッチモジュール
２－０，２－１　ノード制御モジュール
３－１，３－２　通信制御モジュール
４－０－１～４－１－ｋ　ＣＰＵ
５－０，５－１，５－Ｎ　ノード
６－０，６－１　キャッシュモジュール
７－０，７－１　主記憶モジュール
８－０，８－１　入出力モジュール
９－０，９－１，９－Ｎ　ディスク装置
２０　外部クロスバスイッチモジュール
４０　クロスババイパスモード信号ピン
７０－０，７０－１　トランザクション判別部
１００　疑似スヌープ機能付クロスバスイッチ
３００，７１０　クロスババイパスフラグレジスタ
３０３，３０４，３０－２，３０－Ｎ，７０４，７０５　マルチプレクサ
７００　アドレスマップ
８００　要求ノード検索回路

【図１】

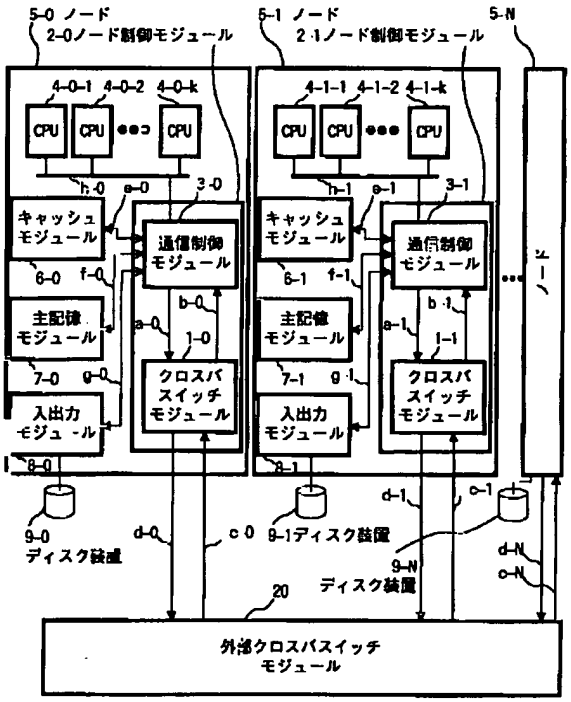


【図６】



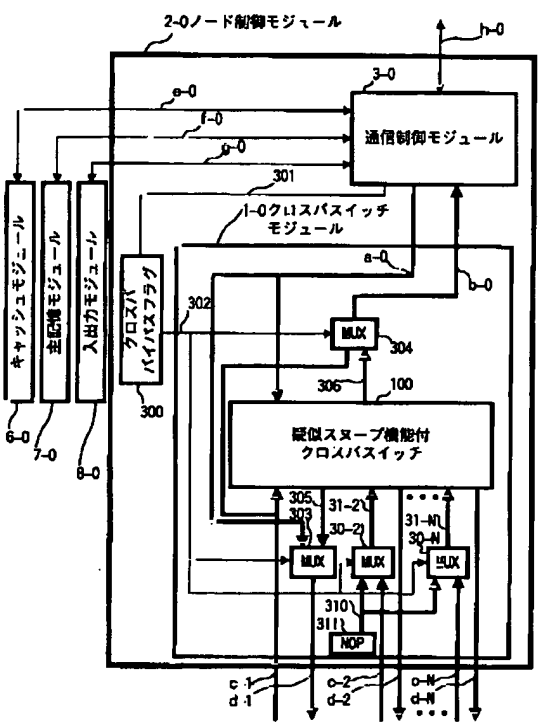
【図2】

図2



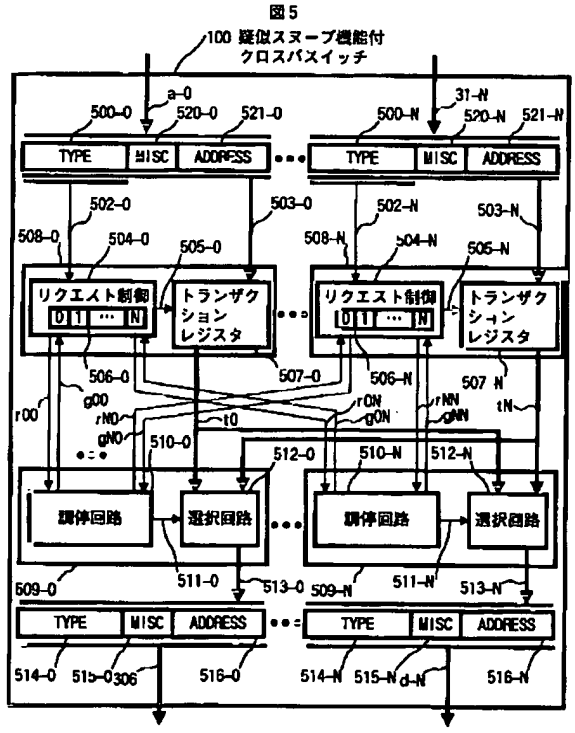
【図3】

図3



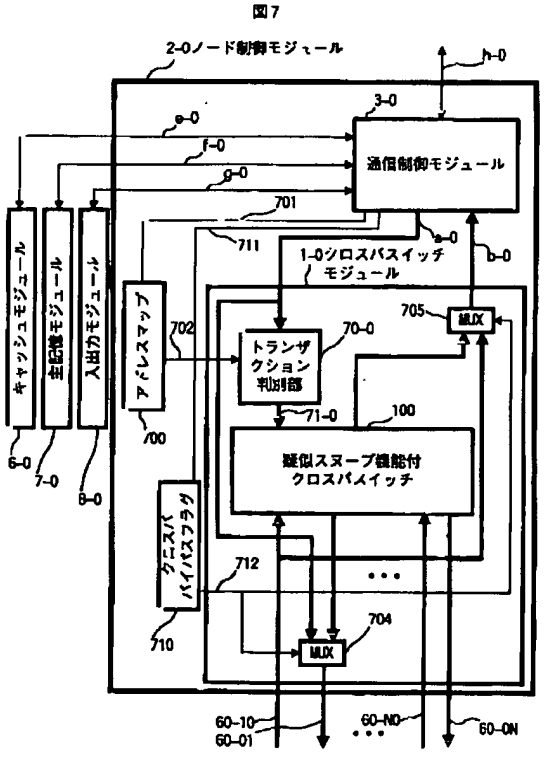
【図5】

図5



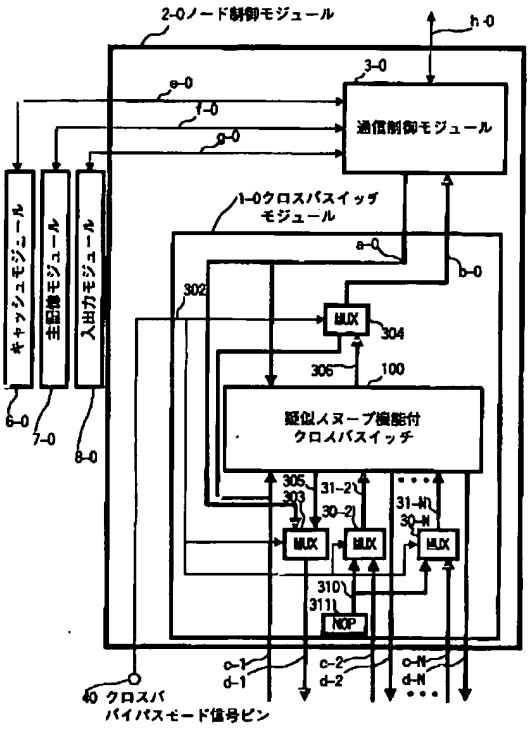
【図7】

図7



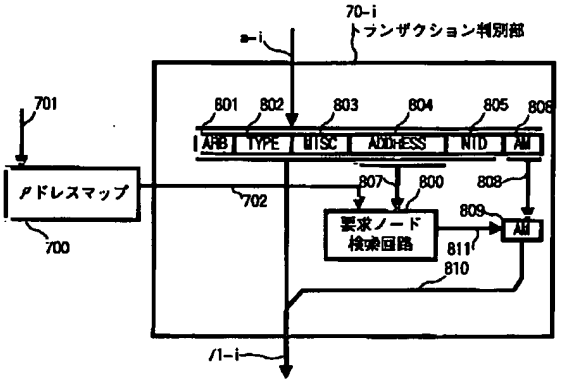
【図4】

図4



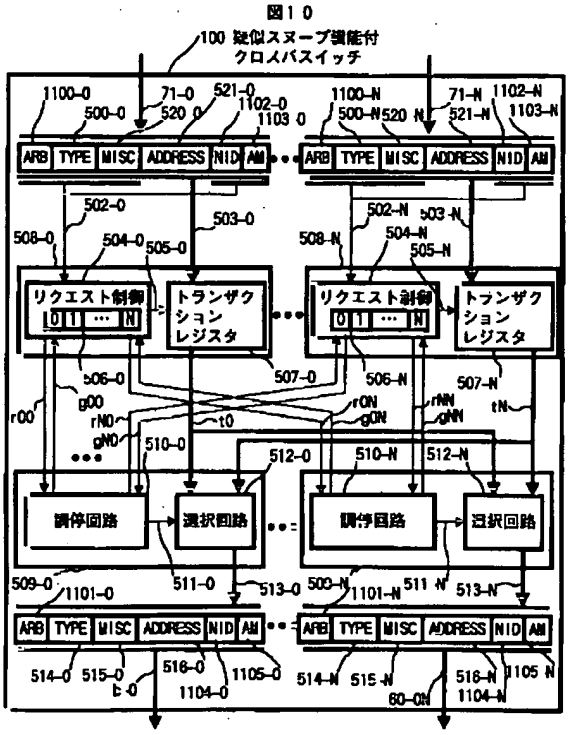
【図8】

図8



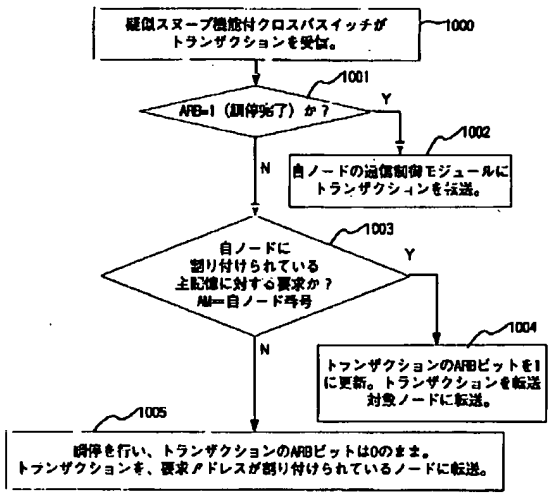
【図10】

図10



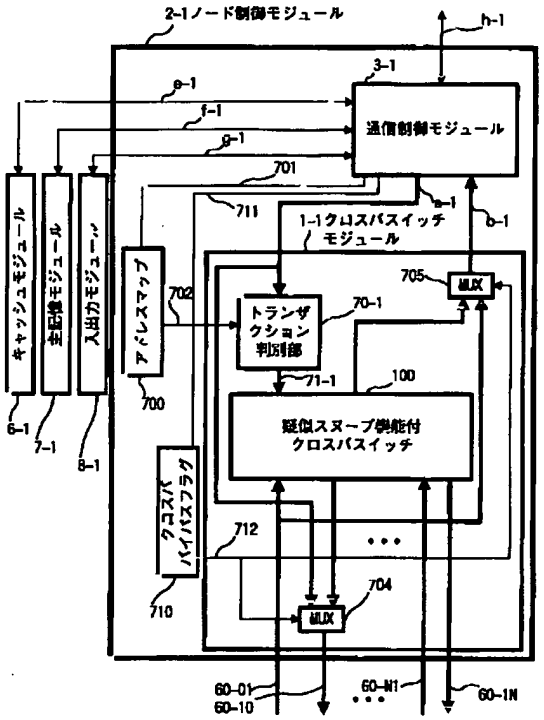
【図9】

図9



【図11】

図11



【図12】

図12

		トランザクションが要求するアドレス空間	
		自ノードの主記憶	他ノードの主記憶
調停前 ARB=0	トランザクション発行元 自ノード	転送対象ノードに転送 ARB=1に更新	要求するアドレスが割り当てられているノードに転送、ARBの更新はしない
	トランザクション発行元 他ノード		
調停後 ARB=1	トランザクション発行 任意のノード	自ノードの通信制御モジュール（最終的には、CPUモジュール、キャッシュモジュール）に転送	

フロントページの続き

(72)発明者 對馬 雄次
東京都国分寺市東恋ケ窪一丁目280番地
株式会社日立製作所中央研究所内
(72)発明者 上原 敬太郎
東京都国分寺市東恋ケ窪一丁目280番地
株式会社日立製作所中央研究所内
(72)発明者 濱中 直樹
東京都国分寺市東恋ケ窪一丁目280番地
株式会社日立製作所中央研究所内
(72)発明者 庄内 亨
東京都国分寺市東恋ケ窪一丁目280番地
株式会社日立製作所中央研究所内

Fターム(参考) 5B045 BB16 BB44 DD12